Clustering short guide:

Print the clustering information to monitor how clusters were formed at different clustering cycles (see Analysis of clustering results):

```
clust -an -p -cy > cluster_formation.out
```

Plot it:

```
plot_cycles.py BnBs cluster_formation.out cluster_formation.png
```

The plot shows how clusters were formed in the last 100 cycles. The red line indicates the minimum distance (1st y axis) between clusters in the last 100 cycles. Blue circles indicate the increase in inter-cluster distance (2nd y axis) for the last 100 cycles. The increase is in % from the most distinct configurations in the pool.

A significant increase in inter-cluster distance is observed at clustering cycle 493, when data is grouped into 7 clusters. To print the information at this level (see Analysis of clustering results):

```
clust -an -p -clcl 7 -f55 fort.55 -p1 p1u.pdb -p2 p2u.pdb > 7_clusters.out
```

Although all configurations are grouped into 7 clusters, only 5 are shown in the output, this is because by default only clusters containing more then 5 members are considered.

To recover the representative configurations of the clusters (see Recovery of the cluster representatives):

```
clust -re -clcl 7 -f55 fort.55 -p1 p1u.pdb -p2 p2u.pdb
```

Recovered cluster representatives are in pdb files cl*.pdb, where * is the number of the cluster.

Clustering program (clust)

Overview

The program can be used to do the clustering of RB docked configurations of protein 2, also scoring, analysis of clustering and recovery of the cluster representatives. It reads directly from the 'complexes' file.

Clustering

Description

The specified number of docked complexes is extracted directly from the 'complexes' file and clustered with hierarchical clustering algorithm. The formation of the clusters and clustering results at every clustering cycle are stored in a binary file (default 'clust.dat').

Parameters

Options	Description
-cl	switch to turn the clustering mode on
-n <int>	number of orientations to take for clustering from file, DEFAULT: 500
-f55 <string>	SDAC output file 'complexes'
-p1 <string>	PDB file of protein 1, same as used in docking
-p2 <string>	PDB file of protein 2, same as used in docking
-clo <string>	output file, DEFAULT: 'clust.dat'

Example

```
./clust -cl -n 500 -f55 complexes -p1 p1.pdb -p2 p2.pdb
```

Clusters the top 500 configurations from the 'complexes' file. The results are stored in the binary 'clust.dat' file (by default) and can be analyzed later with this same program (see below).

Scoring of clusters

Description

When run in scoring mode, the program extracts electrostatic energies and occupancies from the 'complexes' file, calculates backbone RMSD to the starting structure of protein 2, and also calculates the residue propensities for docked configurations. All this information is stored in the binary file 'clust.dat' and is used together with clustering data in the 'clust.dat' file for cluster analysis (see below).

Parameters

Options	Description
-sc	switch to turn the scoring mode on
-f55 <string>	SDA output, file 'complexes'
-p1 <string>	PDB file of protein 1, same as used in docking
-p2 <string>	PDB file of protein 2, same as used in docking
-cdist <float>	contact distance in Å for between residues for residue propensity calculation DEFAULT: 5.0
-rpdef <int>	type of residue propensity calculation DEFAULT: 1
-rpin	a file with residue propensities

Example

```
./clust -sc -f55 fort.55 -p1 p1.pdb -p2 p2.pdb -cdist 5.0 -rpdef 1 -rpin rpFile1 -rpin rpFile2
```

The top 500 configurations from the 'complexes' file are processed

Analysis of clustering results

Description

The formation of clusters at each clustering cycle can be monitored from the stored clustering results. Based on information the user can decide which clustering cycle to use for further analysis. Information about the clusters at the particular clustering cycle can be printed including all scoring data.

Parameters

Options	Description
-an	switch to turn the analysis mode on
-p	switch to turn the printing mode on
-cy <int>	print information about last specified number of clusters DEFAULT: 50 (see Example 1 below)
-bin <int>	print binned representation of the clustering (see Example 2 below)
-clcl <int>	print information on clusters at cycle where specified number of clusters where formed (see Example 3 below)
-clcn <int>	print information on clusters at specified clustering cycle
-clcv <float>	print information on clusters where the smallest distance between clusters is as specified.
-clsz	smallest size of cluster to be printed DEFAULT 5
-f55 <string>	SDAC output, 'complexes' file
-p1 <string>	PDB file of protein 1, same as used in docking
-p2 <string>	PDB file of protein 2, same as used in docking

Example 1 - printing cluster formation data

```
./clust -an -p -cy
```

Tabulated data are printed to the standard output in following format:

Column entry	Description
1	number of clusters
2	clustering cycle number
3	distance between the closest clusters in this clustering cycle
4	the closest distance between clusters increase in % from the previous cycle. The % here is from the whole distance change during the whole clustering
5	the size of the smallest cluster at this cycle
6	the size of the largest cluster at this cycle
7	the average size of the clusters at this cycle

Example 2 - printing binned cluster formation data

```
./clust -an -p -bin 50
```

The closest (smallest RMSD) and the most distinct (largest RMSD) configurations are found in the whole data set. The interval between them is divided into 1000 bins. At every clustering cycle the closest clusters are found and the +1 is added to the bin which has the value of the distance between the closest clusters. In this example the assignment of the last 50 cycles to particular bins is printed out. At the end of the clustering more distinct clusters are formed (larger inter-cluster distance value) and therefore +1 is added to the bins further apart from each other. This shows up as distinct steps when plotted. (plot_bins.py can be used to plot the results)

Example 3 - printing cluster information

```
./clust -an -p -clcl 5 -f55 fort.55 -p1 p1.pdb -p2 p2.pdb
```

Prints out the cluster information to standard output at the clustering level when data is grouped into 5 clusters. (plot_cycles.py can be used to plot the results)

Column entry	Description
1	the number of the cluster (sorted by size)
2	cluster size
3	cluster size with added occurrences from 'complexes' file
4	the number of docked configurations (as ordered in 'complexes' file) which is the representative for this cluster.
5	electrostatic energy of the cluster representative
6	average electrostatic energy in the cluster
7	average distribution of the electrostatic energies in the cluster
8	rmsd of the representative of the cluster relative to starting orientation or protein 2
9	average rmsd of docked orientation relative to starting orientation of protein 2, in the cluster
10	RP score of this cluster (if used)
11	standard deviation of the RP score of this cluster (if used)

Recovery of the cluster representatives

Description

The structures of the cluster representative configurations can be recovered at any clustering cycle as pdb files. The cluster representative is a structure lying in the middle of the cluster

Parameters

Options	Description
-re	switch the recovery mode on
-clcl <int>	recover cluster representatives at the cycle where the specified number of clusters where formed
-f55 <string>	'complexes' file
-p1 <string>	PDB file of protein 1, same as used in docking
-p2 <string>	PDB file of protein 2, same as used in docking

Example

```
./clust -re -clcl 5 -f55 fort.55 -p1 p1.pdb -p2 p2.pdb
```

The cluster representatives will be recovered as pdb files with prefix clX.pdb, where X is the number of the cluster.